시맨틱 웹을 이용한 다국어-지원 신문기사 번역시스템의 설계 및 구현

강정석*, 이기영**
*, ** 을지대학교 의료IT마케팅학과
e-mail:nedved0213@naver.com

Design and Implementation of a Multilingual-Supported Article Translation System using Semantic Web

Jeong-Seok Kang*, Ki-Young Lee**
*, ** Dept. of Medical IT Marketing, Eulji University

요 약

최근 시맨틱 웹의 등장과 발전은 웹 2.0의 발전과 더불어 새로운 웹의 문화를 바꾸어 놓았다. 시맨틱 웹의 적용분야는 다양하지만 그중에서 의미 정보 검색과 다국어 정보 검색 기술을 통한 다국어 지원 번역이 연구 분야로의 필요성이 있다. 기존 기계번역이 번역률에 있어서 가장 큰 한계점은 단어 의미 중의성과 문법적은 오류이다. 따라서 본 논문에서는 시맨틱 웹과 단어 의미 중의성을 해소 시킬 새로운 알고리즘을 제안함으로써 단점을 제거하여 번역률을 향상시켜 모바일에 적용하였다. 모바일에 입력된 신문기사 이미지를 OCR을 통해 텍스트로 변환하고 사전 및 분야 온톨로지와 문장 규칙 추론을 동해 처리 속도 및 정확도 높은 번역시스템을 설계 및 구현하였다.

1. 서론

최근 시맨틱 웹의 등장과 발전은 웹 2.0의 발전과 더불어 새로운 웹의 문화를 바꾸어 놓았다[1]. 또한 웹의 급진적인 발전은 전 세계를 하나로 묶어 정보를 공유하는 사회의 도래를 말한다. 이에 대표적인 매체로는 신문기사가 있다. 세계 곳곳에서 발생하는 사건 및 사고와 해당 나라의 전반적인 상황을 파악하는데 가장 좋은 매체이다. 하지만 임의의 신문기사가 주어졌을 때 그 나라의 언어를 잘 아는 사람 이외에는 어떤 내용인지 쉽게 파악할 수 없는 것이 사실이다. 시맨틱 웹에서의 의미 정보 검색과 다국어 정보 검색 기술은 다국어 지원 번역 연구 분야로의 적용 가능성과 필요성이 있다.

현재 기계 번역의 가장 대표적인 문제점은 단어 의미 중의성과 문법상의 오류이다[2]. 단어 의미 중의성은 같은 단어이지만 여러 개의 의미를 갖는 경우를 말하는데 시소러스의 이용, 정렬 기법, 의미 패턴 등 많은 연구가 이루어졌으나 효과는 미비하다. 문법적인 오류는 다른 어족끼리 긴 문장을 번역시 자주 발생하고 또한 40단어 이상 사용한 문장일 경우 발생 빈도가 높다.

이러한 중의성의 문제나 문법상의 오류는 단어의 의미 때문에 발생하는 문제점이기 때문에 데이터에 의미를 부여하여 사용되는 시맨틱 웹을 통하여 극복 가능하다.

본 논문에서는 모바일 환경에서 신문기사 영상을 텍스트로 변환시키고 언어사전 및 분야별 온톨로지와 SWRL(Semantic Web Rule Language)을 이용한 문장 규칙 추론을 통해 효과적인 번역시스템을 설계 및 구현하였다.

2. 관련 연구

2.1 시맨틱 웹(Semantic Web)

기존 웹에서 HTML을 이용한 표현 방식은 사용자에게 문서 내용을 그대로 보여준다. 따라서 사람이 아닌 기계는 문서로부터 내용의 의미를 알기가 어렵다. 예를 들어, 배라는 단어가 운송수단에서의 배인지, 인체로서의 배인지, 음식으로서의 배인지 구별하기가 힘든 것이다. 이처럼 사용자가 원하는 내용과는 무관한 다른 내용으로 인해서 웹의 효율적인 검색을 방해해왔다[3]. 따라서 데이터에 의미를 부여하여 컴퓨터가 이해할 수 있는 언어로 만들어 컴퓨터에 의해 처리될 수 있도록 제안된 차세대 웹인 시맨틱 웹(Semantic Web)이 연구되었다[1].

컴퓨터가 스스로 웹에 연결된 정보의 의미를 검색하여 그 정보에서 지식을 추론 할 수 있는 기능을 제공한다. 따라서 컴퓨터는 정보의 내용을 이해함으로써 처리 가능하고 또한 추론을 통해서 내재된 지식까지도 처리 가능하다.

시맨틱 웹은 사용자 정의 태그 스키마를 정의할 수 있는 XML과 문서에서 사용되는 의미와 관계를 유연하게 하기 위한 RDF를 바탕으로 구축 된다. 시맨틱 웹에서는 지식에 대한 자원을 유일한 이름으로 지칭하고 온톨로지를 통하여 지식을 공유하는 방식이다[4].

2.2 온톨로지(Ontology)

시맨틱 웹에서는 컴퓨터가 공유하는 데이터들에 대해서 광범위한 개념을 이해하기 위해서는 정형화된 온톨로지에 기반으로 구조적 데이터들로 의존하게 된다. 온톨로지란 정형화된 정보를 개념적이고 컴퓨터에서 다룰 수 있

는 형태로 표현한 모델로 개념의 타입이나 제약조건들을 명시적으로 정의한 기술이다. 특정 도메인 내에 정의된 지식을 바탕으로 개념화 및 명세화 함으로써 시스템 간 정보의 공유와 재사용이 가능하다[4]. 앞에서 언급하다시피 시맨틱 웹에서는 온톨로지를 기반으로 데이터의 공유가 이루어지므로 온톨로지에서 사용한 용어들은 기본적인 개념과 정의를 통해 선택되어야하고 또한 높은 지식수준을 제공해야한다.

2.3 OCR(Optical Character Recognition)

OCR이란 사람이 쓰거나 기계로 인쇄한 문자의 영상을 이미지 스캐너로 획득하여 기계가 읽을 수 있는 문자로 변환하는 것이다[5]. 이미지 스캔으로 얻을 수 있는 문서의 활자 영상을 컴퓨터가 편집 가능한 문자 형식으로 변환하여 인식하는 방식이다.

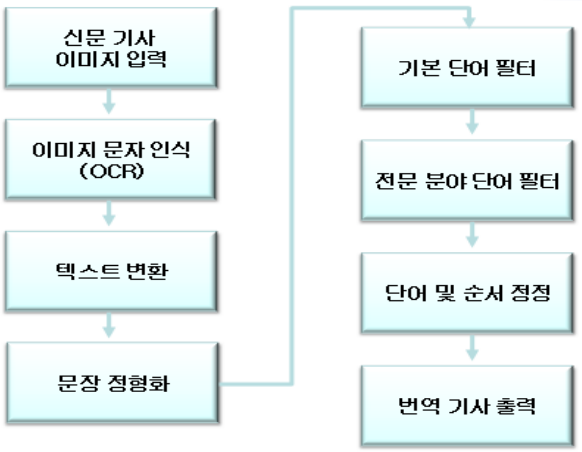
OCR을 하기 위해서는 먼저 컬러 영상을 회색조 영상으로 만들어주는 이진화 과정을 수행한다. 이것은 모바일이 영상을 디지털 정보로 인식하게 만드는 기술로 OCR에서는 핵심적인 기술 중에 하나이다.

처음 이미지를 입력 받으면 대부분의 이미지는 OCR 인식 하기 어려운 다양한 형태의 사각형의 모양이다. 따라서 다양한 사각형을 인식하기 쉬운 직사각형 모양으로 변환하는 기울기를 보정하는 과정을 수행해야 한다. 이는 근접 문자들 간의 관계를 규명함으로써 사각형의 꼭지점 4개를 찾아 직사각형을 만들어 내는 과정으로 이루어진다.

그 후 영상에 있는 문자를 추출해야 하는데 문자 추출하는 방식은 먼저 영상을 x,y축 방향으로 프로젝션을 적용하여 검은 픽셀이 존재하는지 검사 후 문자를 추출 한다. 그리고 마지막으로 문자를 인식하는 과정을 수행한다. 이 과정은 문자 추출 과정에서 생긴 특징데이터를 표준문자의 특징 데이터와 비교하여 근사치에 해당하는 값을 해당 단어로 인식하게 된다[6].

3. 시스템 설계

3.1 시스템 흐름

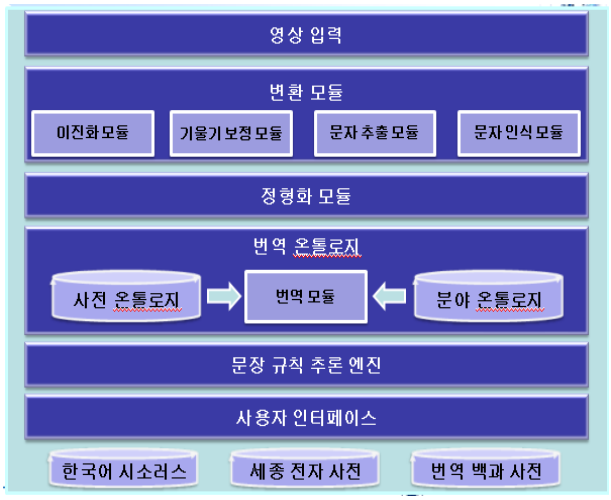


(그림 1) 시스템 흐름도

본 논문에서의 시스템은 신문기사의 이미지를 OCR인식을 이용하여 텍스트로 변환하고 정형화를 통해 각 단어의 의미와 의미 관계를 파악한다. 이를 바탕으로 구축된 언어 정보 및 분야 온톨로지와 문장 규칙 추론을 통해 번역이 이루어지는 시스템이다.

또한 제안한 시스템에서는 높은 번역률을 얻기 위해서 단어 의미 중의성을 해소할 수 있는 효과적인 알고리즘을 제시하여 기계 번역시 가장 큰 단점인 단어 의미 중의성과 문법상의 오류를 줄일 수 있다. 제안한 시스템의 순서도는 그림 1과 같다.

3.2 시스템 구성



(그림 2) 시스템 구조도

본 논문에서 제안한 시스템의 전체 구성은 그림 2와 같은 구조로 이루어져 있다. 시스템을 수행하기 위해서 먼저 신문기사의 영상을 입력해야 한다. 이 때 입력된 영상은 10~30cm의 제한 거리가 적용 되어야 한다.

변환 모듈에서는 입력받은 영상의 신문기사를 OCR인식 과정을 통해 문자로 변환하는 과정으로 이진화 모듈, 기울기 보정 모듈, 문자 추출 모듈, 문자 인식 모듈로 구성 되어 있다. 이진화 모듈에서는 컬러 영상을 회색조 영상으로 만들어서 영상을 디지털 정보로 인식하게 만들고, 기울기 보정 모듈에서 영상을 직사각형 모양으로 기울기를 보정하여 인식 과정을 용이하게 한다. 문자 추출 모듈과 문자 인식 모듈에서 최종적으로 영상을 문자로 변환해주는 작업을 수행하는데 검은 픽셀의 존재 유무를 검사하여 문자를 추출하고 특징 데이터와의 비교로 문자를 인식하게 된다.

이렇게 변환된 문자들을 일괄적으로 처리가능 하도록 단어들의 격, 의미 관계 및 문장 순서 등을 파악하고 변형하는 정형화 모듈을 거치게 된다.

파악된 정보들을 통하여 번역이 이루어지는데 번역 모듈은 한국어 시소러스, 세종 전자 사전, 번역 백과사전을 기반으로 하여 구성되어있다. 번역의 과정은 2단계로 이루어진다. 먼저 사전 온톨로지를 통하여 기본 단어들을 필터

링 한다. 그 후 신문기사 분야의 온톨로지를 통해 필터링을 하여 번역이 이루어진다. 이로 인해 처리 속도 및 정확성은 향상 될 수 있다. 또한 번역률을 높이기 위하여 단어 의미 중의성을 해소 시킬 수 있는 그림 3과 같이 새로운 알고리즘을 제안하였다. 먼저 사전 온톨로지에서 표제어 테이블을 추출하고 형태소 정보를 비교한다. 물론 동일하면 그대로 결과를 출력하면 되지만 동일하지 않을 경우 단어 의미 수의 개수를 확인하고 하나 이상이면 격 정보와 선택 제약 정보를 비교 후 같으면 해당 의미로 결정, 다르면 최다 빈도수 의미로 결정된다.

1 :	표제어 테이블 추출 후 비교
2 :	단어 의미 수 확인
3 :	하나 이상이면 격 정보와 선택 제약 정보 비교
4 :	같으면 해당 의미로 결정
5 :	다르면 최다 빈도수 의미로 결정

(그림 3) 단어 의미 중의성 알고리즘

번역이 완성된 문장은 문장 규칙 추론 엔진을 통하여 번역 된 언어에 맞는 문장 규칙 및 순서를 재배열 하게 된다. 문장 규칙 추론 엔진은 시맨틱 웹에서 규칙은 정의하는 언어인 SWRL로 구축한다. 이러한 과정을 통하여 신문기사 이미지를 문자로 변환 타 언어로 번역하는 과정을 수행하게 되는 시스템이다.

4. 성능 평가

기존에 많이 사용되고 있는 기계 번역에서 높은 번역률을 얻을 수 없는 이유는 단어 의미 중의성과 문법적인 오류이다. 단어 의미 중의성은 여러 가지의 의미를 가지고 있는 단어를 사용했을 때 문장 전체의 의미가 아닌 기계가 경험적으로 가장 많이 사용한 의미를 번역에 적용하므로 번역률이 떨어지고 문법적인 오류는 통상 40단어 이상의 단어로 이루어진 문장을 번역 시 해당 언어와 번역하고자 하는 언어가 다른 어족일 경우 문법적으로 다르기 때문에 오류가 발생한다.

이 두 가지의 단점을 보완하기 위하여 시맨틱 웹을 사용하고 새로운 단어 의미 중의성 알고리즘은 제시하였다. 시맨틱 웹을 사용함으로써 사전 온톨로지를 통해 조회하기 때문에 처리속도 및 단어 의미 파악을 높일 수 있고, 문장 규칙 추론 엔진을 통하여 문장에서 사용 된 A라는 단어가 여러 가지 의미가 있을 경우 B라는 단어와 같이 쓰일 때는 이런 의미로 번역 하게 된다. 그리고 새로운 단어 의미 중의성 알고리즘을 제시함으로써 중의성 문제도 잠식시킬 수 있다.

5. 결론

기계 번역은 많이 사용되고 있지만 번역률이 저조할 수밖에 없는 결정적인 단점을 보유하고 있는 실정에서 최근 웹 2.0의 발전으로 인하여 시맨틱 웹의 등장과 급진적인 발전이 이루어졌다. 그리고 시맨틱 웹에서는 다국어 지원 번역 기술이 연구 분야로서 적용 가능성과 필요성을 가지고 있기 때문에 기존 번역 기술에 시맨틱 웹을 적용하였다.

따라서 본 논문에서는 OCR기술을 이용하여 신문기사 이미지를 텍스트로 변환 시키고, 기계 번역에 시맨틱 웹을 적용하여 설계 및 구현하였고, 새로운 단어 의미 중의성 알고리즘을 제시하여 처리 속도 및 번역률을 향상시켰다.

향후에는 OCR 기술 중 문자 추출과 문자 인식에 관해서 인식율 높이기 위해 새로운 알고리즘을 연구할 것이다. 또한 신문기사의 한 분야가 아닌 여러 파트의 온톨로지를 구축하여 신문에서 어느 분야의 이미지를 얻을 경우 인식율 및 번역률을 높일 수 있게 새로운 영역으로 확장시킬 것이다.

참고문헌

[1] 한성국, 정영식, 유재국, “웹2.0과 시맨틱웹, 그리고 진화의 방향”, 한국정보과학회 정보과학회지, 제25권 제10호, 57~66쪽, 2007년

[2] 최승권, 홍문표, 박상규, “다국어 자동번역 기술”, 한국전자통신연구원 전자통신동향분석, 제20권 제5호, 16~27쪽, 2005년

[3] Danushka Bolegala, Yutaka Matuso, Mitsuru Ishizuka, ″Measuring Semantic Similarity between Words Using Web Search Engines″, International World Wide Web Conference, pp. 757~766, 2007

[4] 김현주, 최중민, “온톨로지 생성과 공유를 위한 시맨틱 웹 기반 위키 시스템”, 한국정보과학회 정보과학회논문지 : 소프트웨어 및 응용, 제33권 제8호, 703~717쪽, 2006년

[5] P.M. Hall, ″The Practical Optical Character Recognition System″, Electronics and Power, vol. 14, pp. 149~153, 2009년

[6] 박종경, 음봉규, 권용식, 진성아, “모바일 환경의 OCR Anyword”, 한국콘텐츠학회 한국콘텐츠학회 2006 춘계종합학술대회논문집, 제4권 제1호, 152~155쪽, 2006년